International Core Journal of Engineering Volume 6 Issue 3, 2020
ISSN: 2414-1895 DOI: 10.6919/ICJE.202003_6(3).0048

Research on UAV Flight Control Base on Speech Recognition

Yuanliang Zhang
School of Shanghai Maritime University, Shanghai 201306, China.
yuanliang_zh@163.com

Abstract

Research on intelligent voice control of drones needs to realize accurate and efficient recognition of operator voice commands by drones. However, traditional Dynamic Time Warping(DTW) algorithms need to store large matrices, which require a large amount of calculation, high time complexity, and low algorithm recognition rate when directly calculating. In view of these problems, this paper proposes an optimization algorithm for polygonal curved window constraints. First, use Mel-scale Frequency Cepstral Coefficients to extract the voice command feature information. Then, based on the Itakura Parallelogram diamond curved window constraint DTW optimization algorithm, a DTW optimized matching algorithm for polygon curved window constraints is further proposed. Finally, the optimization coefficients are used to achieve the optimal polygon constraint. DTW algorithm scheme. Experimental results show that the optimal algorithm is robust and time-effective, and can be used as a command input port for intelligent control of drones.

Keywords

Speech recognition; MFCC; Dynamic time warping; Curved window.

1. Introduction

At present, speech recognition technology is widely used in various fields such as computers, robots, smart homes, and mobile phones. This technology makes human-computer interaction more convenient, fast, and efficient. However, voice control technology is rarely used in the field of drones, and it mainly depends on the operator to use the remote control and other equipment for on-site control, which has certain requirements for the operator [1].

In order to reduce the requirements for drone operators and operation errors caused by drone crashes, while referring to the application research of speech recognition technology in the field of autonomous driving, this article proposes a voice control scheme for the drone control field.

UAV voice control belongs to isolated word recognition, which requires high accuracy and fast recognition speed. Although there are many efficient algorithms for phrase sound recognition, such as GMM-HMM, DNN-HMM, and full convolutional neural networks [2]. However, the accuracy of these algorithms depends on the training of a large number of pronunciation samples, and they are relatively complicated and computationally intensive, so they are more suitable for continuous speech recognition with large vocabulary [3]. In contrast, DTW technology can learn and train with less sample data, and can get better speech models, which is more suitable for isolated word recognition in the absence of samples. Therefore, in this study, DTW algorithm is used to recognize the drone voice instructions. However, the DTW algorithm recognizes speech by expanding several path searches in the entire rectangular curved window to find the best matching path [4]. The time complexity and space complexity are relatively high, so the algorithm has low recognition efficiency.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202003_6(3).0048

In view of the above-mentioned problem of low recognition efficiency of the DTW algorithm, there are currently two main improvement methods. The first is the early termination technique: this method stops searching when the cumulative bending cost exceeds a certain threshold, and considers that the two speech sequences do not match. This method can effectively reduce the computational complexity of the algorithm. The second is an elastic coarse-grained dynamic bending time series similarity algorithm: the idea is to reduce the dimensionality of time series data through data processing and replace the original time series with low-dimensional features, thereby greatly improving the DTW calculation efficiency. The disadvantages of the above two improved methods are to improve the efficiency of the algorithm at the cost of losing the recognition accuracy of the algorithm.

Due to the high safety requirements of drone control, the existing improved DTW algorithms mentioned above are difficult to meet the high performance requirements of drone control. In order to reduce the complexity of the algorithm while retaining a certain degree of accuracy, this paper proposes a polygon-constrained DTW algorithm combined with MFCC feature extraction. First, use MFCC to extract features from speech signals, and then use these feature data to perform global DTW search [5]. Aiming at the low recognition rate of DTW algorithm, this paper proposes a global optimization algorithm for curved windows with hexagonal constraints based on the DTW algorithm of Itakura Parallelogram.

The first part of this article first introduced the overall design of the system and the implementation process. The process of feature extraction is described in the second part. The third part of the article describes the DTW search principle and proposes a polygon window constraint optimization algorithm. In the fourth part, it is verified through actual experiments that this scheme improves the accuracy of the algorithm.

2. System Overview

The drone voice control system designed in this paper is an intelligent system that indirectly controls drones through specific human voices, as shown in Fig 1. The system's work flow is: pre-train the operator's voice command model and save the trained command feature data. Then use a mobile phone to collect the voice signal and perform recognition processing on the voice signal. Finally, according to the obtained results, it is mapped into corresponding drone flight instructions, which are transmitted to the drone through Wifi, and the drone flight is controlled indirectly.

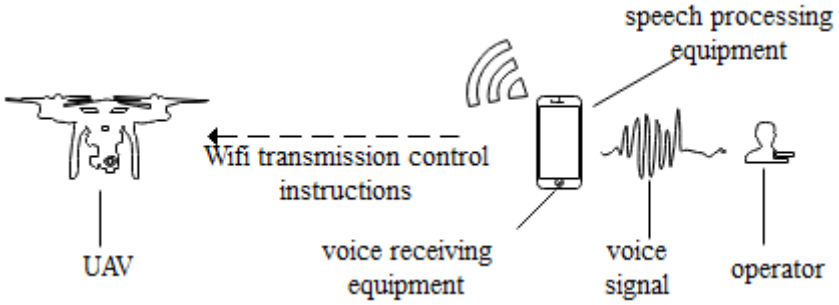

Figure 1. Schematic diagram of drone voice control system

Speech recognition can be divided into two processes: speech template library training and speech signal matching, as shown in Figure 2. Regardless of training or recognition, the speech signal needs to be pre-processed. The process mainly includes signal filtering, endpoint detection and feature extraction. Usually, the recorded sound signals are inevitably noisy. In order to prevent noise from affecting the endpoint detection of sound signals and the extraction of feature vectors, noise reduction must be performed before processing the signals [6]. In this paper, the minimum mean square error estimation based on the short-term book spectrum is used to perform speech filtering. In order to further improve the accuracy of the system, it is also necessary to perform voice activation detection

ISSN: 2414-1895 DOI: 10.6919/ICJE.202003_6(3).0048

on the voice signal and cut off the mute section in the signal. This operation of mute resection is generally called VAD, and a double threshold detection method is used in this paper.

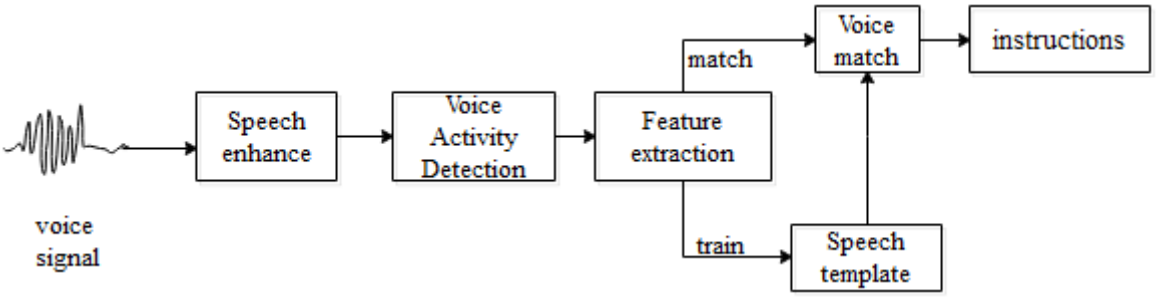

Figure 2. Voice command recognition process

3. Feature Extraction

In speech recognition, commonly used speech feature extraction includes Linear Predictive Coding (LPC) and Mel-scale Frequency Cepstral Coefficients (MFCC). The MFCC takes into account the human auditory characteristics, first maps the linear spectrum to the Mel nonlinear spectrum based on auditory perception, and then converts it to the cepstrum. In the Mel frequency domain, people's perception of pitch is linear. The nature of a Mel filter is actually a rule of scale, usually a triangle filter bank that passes energy through a pair of Mel scales [7]. As shown in Figure 3, a filter bank with M filters is defined, the filter used is a triangular filter, the center frequency is, and M is usually taken from 22-26. The interval between f(m) is reduced by shifting the value of m, and is increased by increasing the value of m.

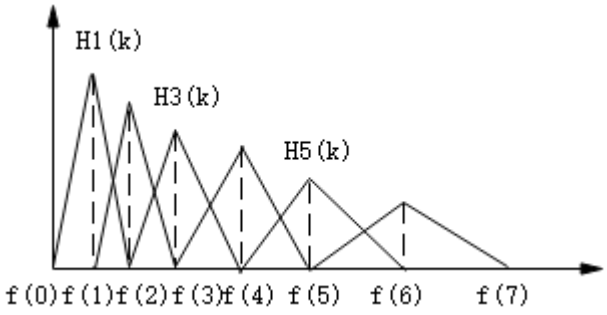

Figure 3. Mel filter

MFCC is a cepstrum parameter extracted in the frequency domain of the Mel scale. The Mel scale describes the non-linear characteristics of the frequency of the human ear. Its relationship with frequency can be approximated by the following formula:

$$\mathrm{Mel}(f) = 2595 \times \lg(1 + \frac{f}{700}) \quad (1)$$

MFCC speech feature extraction process, as shown in Figure 4. Pre-emphasis, framing, and windowing need to be performed on the speech signal, and these processing methods are designed to maximize certain information of the speech signal to achieve the best feature parameter extraction.

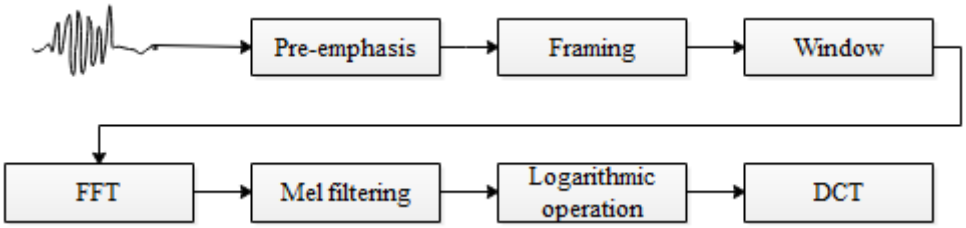

Figure 4. MFCC speech feature extraction

ISSN: 2414-1895 DOI: 10.6919/ICJE.202003_6(3).0048

4. DTW Dynamic Recognition Algorithm and Optimization

4.1 Basic principles of traditional DTW algorithm

In isolated word speech recognition, standard speech is trained to generate a standard speech template feature library. After the recognized speech undergoes the same filtering, endpoint detection, and feature extraction, a test template is obtained. The main process of identification is the matching process between the test template and the standard template. The matching process is to find the similarity between the two, and the similarity can be described by the distortion between the two. The smaller the distortion, the higher the similarity.

Assume that the standard template speech time series is Q, $Q = q_1, q_2, \cdots, q_N$, and the test template speech time series is C, $C = c_1, c_2, \cdots, c_M$. Through these two time series, a matrix $N \times M$ grid is constructed. The matrix elements (i, j), represent the distance $d(q_i, c_j)$ between the two points q_i and c_j, which is also called the degree of distortion.

Each point in the matrix constructed by the sequences Q and C has the possibility of forming a curve. The DTW algorithm is to find a shortest path that meets the conditions. The search path of the DTW algorithm is shown in Figure 5.

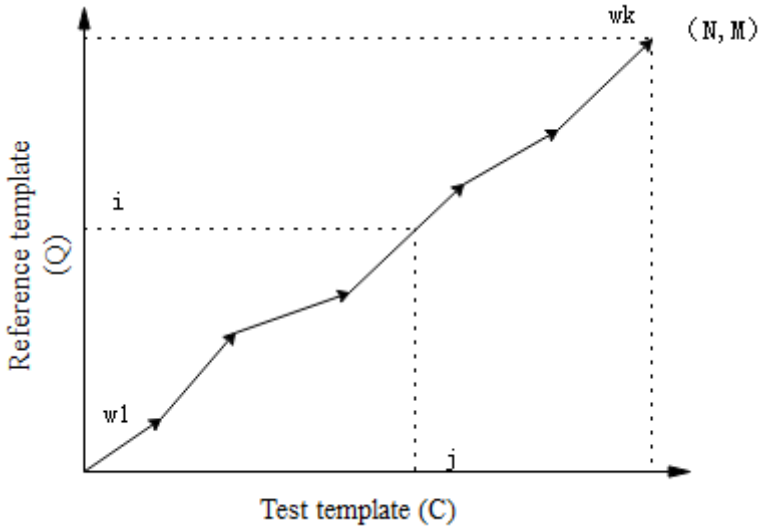

Figure 5. DTW search path

Define this path as a regular path and use W to represent it, and the k-th element of W is defined as $w_k = (i, j)_k$, which represents the mapping of sequences Q and C,

$$W = w_1, w_2, \cdots, w_k \tag{2}$$

At the same time, the path must satisfy three conditions: boundary conditions; continuity; and monotonicity.

Then the smallest regular path can be described as

$$D(Q,C) = \min_c \left(\frac{\sum_{k=1}^{K} W_k}{K}\right) \tag{3}$$

The K in the denominator is mainly used to compensate for regular paths of different lengths.

4.2 DTW algorithm optimization

The traditional DTW algorithm starts from (0, 0) and ends at (N, M). It expands several search paths in the entire rectangle and calculates the accumulated distance corresponding to the matched speech frame. The one with the smallest accumulated distance is the best search path. This algorithm is computationally intensive and its time complexity is O (MN).

ISSN: 2414-1895 DOI: 10.6919/ICJE.202003_6(3).0048

Aiming at the problem that the algorithm is computationally intensive, Itakura proposed an Itakura Parallelogram global constraint window, whose window is shown in Figure 6.

Where the slope of the diamond window is 0.5 or 2, and the coordinates of the four vertices are (0, 0), (N, M), $((2M-N)/3,(4M-2N)/3)$ and $((4N-2M)/3,(2N-M)/3)$, the expressions for the four edges are $y=2x$, $y=0.5x$, $y=0.5(x-N)+M$ and $y=2(x-N)+M$.

By adding a diamond window, the DTW search path is limited to a diamond window, thereby reducing the calculation amount of the algorithm. This article further adds constraints on the basis of the diamond window, as shown in Figure 7.

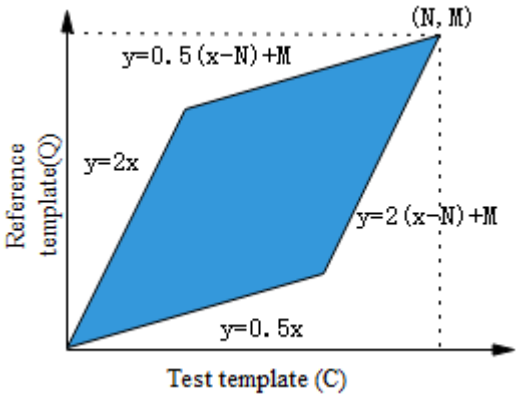

Figure 6. Itakura Parallelogram global constraint window

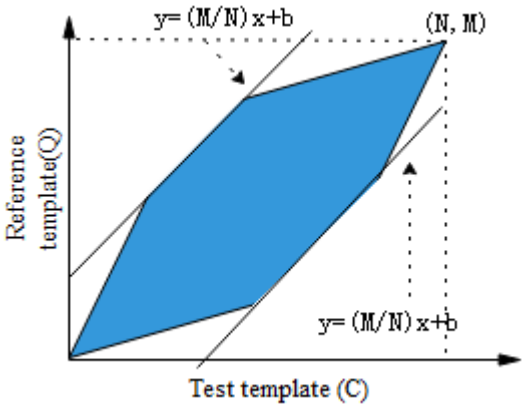

Figure 7. Improved constraint window

Adjust the slope of the diamond window to k and add two straight lines $y=(M/N)x+b$ and $y=(M/N)x-b$ to make it a hexagon window. The six edges of the window can be expressed as:

$$\begin{cases} y=kx \\ y=(1/k)x \\ y=(M/N)x+b \\ y=(M/N)x-b \\ y=(1/k)(x-N)+M \\ y=k(x-N)+M \end{cases} \tag{4}$$

Where, $1<k<\infty$ and $0<b<\dfrac{MNk^2-(M^2+N^2)k+MN}{N(k-1)}$, vertex coordinates are $(0,0)$, (N,M), $(\dfrac{Nb}{Nk-M},\dfrac{Nbk}{Nk-M})$, $(\dfrac{Nbk}{Mk-N},\dfrac{Nb}{Mk-N})$, $(\dfrac{(M-b)Nk-N^2}{Mk-N},\dfrac{(M-b)Mk-MN}{Mk-N}+b)$, and $(\dfrac{Nk^2-(M+b)N}{Nk-M},\dfrac{Mk^2-(M+b)M}{Nk-M}-b)$.

It can be known from formula (4) that the hexagon window is affected by the slope of the diamond k and the intercept b of the line. Therefore, the global optimum can be achieved by adjusting the slope k and the intercept b.

5. Experimental results and analysis

In order to obtain voice data, seven basic instructions were recorded in this experiment using a Windows recording device, including "takeoff, landing, forward, backward, left, and right," with a sampling frequency of 44.1kHz and a sampling number of 16bit. At the same time, set the acquisition time of each instruction to 2s.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202003_6(3).0048

The voice data is further divided into a reference template and a test template. The training sample collects voice data according to the above environment, collects 100 times for each instruction, extracts MFCC features, and resumes speech templates for 100 samples of each instruction. Seven instructions with a total of 700 reference templates constitute the isolated word instruction speech library of this study. The test template is established in the same way as the reference template, and 700 test templates are established to form a test voice set.

Match each template data in the test set with the standard template library, first adjust the rhombus slope k, where k is from 1 to 10, and take a value every 0.5 interval. The result is shown in Figure 8.

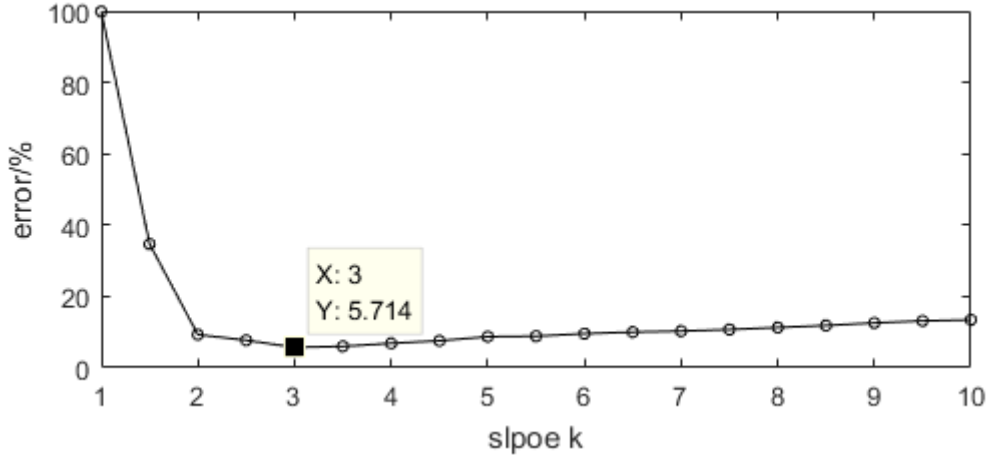

Figure 8. The recognition error rate of the algorithm at different slopes k

From the analysis in fig. 8, it can be known that: At this time, the recognition accuracy of the curved window is from low to high, and then from high to low, and the optimum is achieved when the slope is k = 3.

In the case of slope k = 3, in order to ensure that the straight line and the rhombus have two intersections, it must ensure $0 < b < \frac{10MN - 3(M^2 + N^2)}{2N}$. Each time b is scaled by a different multiple of the boundary, the result is shown in Figure 9.

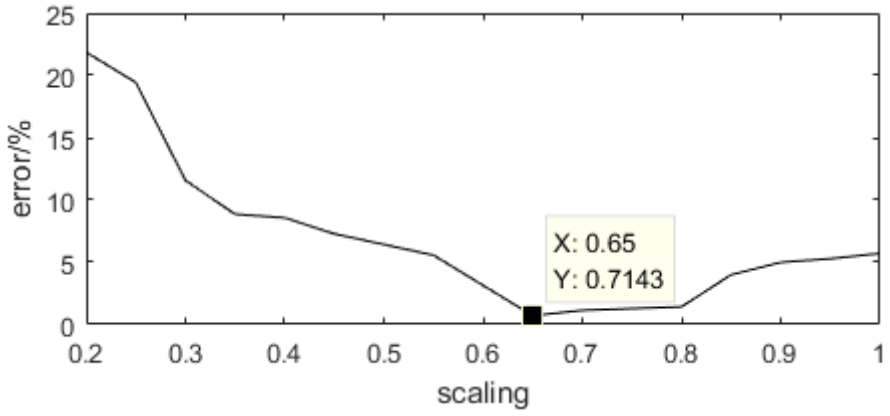

Figure 9. Algorithm recognition error rate at k = 3 and different intercepts b

It can be seen from the analysis in figure 9 that when k = 3, as the intercept b increases, the error rate of the algorithm decreases first and then increases.

6. Conclusion

The DTW algorithm will expand several search paths in the constraint window, and its calculation volume and accuracy will change with the size of the window. The experiments show that the improved polygon window can adjust the parameters appropriately, so that the DTW algorithm under the constraints of the window can meet the requirements of the UAV intelligent voice control system.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202003_6(3).0048

References

[1] XU M, WANG B, WANG K. Discussion of cockpit voice instruction design. Avionics Technology, 2012, No. 43(3), p.39-43.

[2] CAI Q L, CHEN L, SUN J L. Piecewise statistic approximation based similarity measure for time series. Knowledge-Based Systems, 2015, No. 85, p.181-195.

[3] Research of Remote Measurement and Control Technology of UAV Based on Mobile Communication Networks. International Conference on Information and Automation, Lijiang, China. 2015 No. 472(4) p.2517-2522.

[4] Afonso L, Souto N, Sebastiao P, et al. Cellular for the skies: Exploiting mobile network infrastructure for low altitude air-to-ground communications. IEEE Aerospace and Electronic Systems Magazine, 2016, No. 31(8) p.4-11.

[5] Manuel J. Reinoso, Luis I. Minchala, Paul Ortiz, Darwin F. Astudillo, and Diego Verdugo ”Trajectory tracking of a quadrotor using sliding mode control”, IEEE Latin AmericaTransactions, p. 2157-2166, August 2016.

[6] Samir Zeghlache and Abderrahmen Bouguerra, “Sliding mode control based on interval type-2 fuzzy-neural network controller for an UAV”, 2017 10th International Conference on Electrical and Electronics Engineering (ELECO), pp.780-783, January 2018.

[7] Fuyang Chen, Rongqiang Jiang, Kangkang Zhang, Bin Jiang, and Gang Tao, “Robust Backstepping Sliding-Mode Control and Observer-Based Fault Estimation for a Quadrotor UAV”, IEEE Transactions on Industrial Electronics, pp.5044 – 5056, April 2016.